User Fairness in NOMA-HetNet Using Optimized Power Allocation and Time Slotting

User Fairness in NOMA-HetNet using Optimized Power Allocation and Time Slotting

Pragya Swami, Vimal Bhatia, *Senior Member, IEEE*, Satyanarayana Vuppala, *Member, IEEE*, and Tharmalingam Ratnarajah, *Senior Member, IEEE*

***Abstract*—Exponential growth in number of users with diverse data rate requirements has lead to the heterogeneity of traditional cellular networks. To support massive number of users, non-orthogonal multiple access (NOMA) has emerged as a promising solution to achieve increased number of connections and higher spectral gains as compared to orthogonal multiple access (OMA). However, studies show that weak users (WU) and strong users (SU) served using NOMA (referred as NOMA-group) experience different throughputs. In a NOMA group, an SU achieves higher throughput than a WU. Further, as the number of users in a NOMA group increases, due to superposition of signal of multiple users in NOMA, the intra-group interference dominates, thereby reducing throughput of the WUs. This work proposes novel time slotting (TS) techniques that aims at user fairness amongst the users by increasing the throughput of WUs, especially when the number of users increases in a NOMA group. The power allocation coefficients and the time slot duration for the proposed TS techniques are optimized to satisfy the minimum throughput of each user in a NOMA group while maximizing the throughput of WUs. The fairness between the users is measured by calculating both quality of service fairness and quality of experience fairness experienced by the user. It is observed that the proposed TS technique improves the fairness measures significantly. Furthermore, energy efficiency (EE) is also calculated for the TS techniques using the optimized power allocation coefficients and time duration. The numerical results suggest improvement in the EE of the system along with enhancing user fairness amongst the users.**

***Index Terms*—Non-orthogonal multiple access, use fairness, fairness index, heterogeneous network, throughput, energy efficiency.**

I. Introduction

To serve large number of users with diverse requirements, non-orthogonal multiple access (NOMA) [1] is a viable solution for future wireless networks. Power Domain (PD) NOMA attains multiplexing in power domain by assigning different power allocation coefficients to different users served using PD-NOMA (hereafter referred as NOMA group). Performance of PD-NOMA is studied in [1], [2], which proves its superior throughput. The importance of selecting appropriate power allocation coefficients in NOMA to outperform the conventional orthogonal multiple access (OMA) technique is studied in [3]. To differentiate the users served using NOMA, in this work, we classify the users in two categories based on their channel condition; namely weak users (WUs) and strong users (SUs). Assuming perfect knowledge of users' channel state information (CSI) at base station (BS), the WU is defined as a user with poor channel condition, for instance, user in the cell edge region. The SU is defined as a user with good channel condition, e.g., users in the cell center region. The BS pairs/groups[1] users with different channel condition and serves them using NOMA. The authors in [4], [5] prove that in a NOMA group a user with better channel condition, i.e., an SU achieves higher throughput in comparison to a user with poorer channel condition, i.e., a WU.

Due to the degraded rate achieved by the WUs, to maximize the sum rate of the system, the WU's are not favored for resource allocation. Hence, their performance is substantially compromised causing unfairness amongst the SU and WU [6]. The trade-off between sum rate performance and fairness can be analyzed and balanced using a metric that measures fairness amongst the users. The fairness can be based on either quality of service (QoS) or quality of experience (QoE) perceived by the user. The commonly used QoS fairness metric in the literature is Jain's fairness index [7], [8] and is extensively used in wireless networks with NOMA to balance user fairness and network sum rate. While QoS fairness has been well studied, focus on fairness from perspective of the users need to be established. The work in [9], [10] argues that a it is not necessary that a system which is QoS fair is also QoE fair. Hence, it is important to consider fairness from the QoE perspective as well. QoE is evaluation of media quality at individual users. Mean Opinion Score (MOS) is one of the commonly used evaluation methods to characterize the QoE experienced by users [10], [11].

Furthermore, to eliminate unfairness in NOMA networks, majorly three strategies are used. In the first strategy cooperative NOMA is used wherein a nearby user (i.e., an SU) is treated as a relay to assist a distant user (i.e., a WU) as studied in [12]. The authors in [13] use energy harvesting at the SU to assist the WUs using cooperative NOMA. The second strategy is to add more design variables for fairness amongst the users, e.g., weighted sum-rate [14], [15]. The third strategy is to enhance performance of the WUs as studied in [16], [17] while maintaining minimum requirement of the SU.

Furthermore, rapidly increased users and huge data demands has lead to the conventional network comprising of only

This work is an outcome of the R&D work undertaken under the Visvesvaraya PhD Scheme of Ministry of Electronics and Information Technology, Government of India, being implemented by Digital India Corporation and also in part by DST UKIERI (DST/INT/UK/P-129/2016 and DST UKIERI-2016-17-0060).

P. Swami and V. Bhatia are with Indian Institute of Technology Indore, Indore, India; S. Vuppala is with United Technologies Research Center, Cork, Ireland, and T. Ratnarajah is with the Institute for Digital Communications, University of Edinburgh, Edinburgh, UK.

[1]The main focus of the proposed work is not on how the CSI is acquired by the BS or how the pairing/grouping is done by the BS. Rather, the main contribution of the proposed work is to enhance performance of the WUs by reducing the intra-group interference in the group formed at the BS.

macro base station (MBS) tier to shift towards more practical heterogeneous cellular networks (HetNets) [18], [19]. The HetNets comprises of the MBS tier deployed with small base stations, e.g., femto base station (FBS) tier, to aid the MBS tier, especially in the overcrowded areas such as shopping malls, sports venues, airports and others. Offloading in the HetNets plays a viable role in the load balancing by handing some users to the FBS tier and is studied in detail in [18], [20]. In this work, assuming open access FBS [21], when the MBS tier is congested, macro users (MU) can be offloaded to the FBS tier. NOMA is employed in the HetNet [22], at the FBS tier, and the offloaded MU is paired with the available femto users (FU) at the FBS tier and served using NOMA [18]. Authors in [18] performs offloading from the MBS tier to the FBS tier. The offloaded MU is paired with an FU and the two users are served using NOMA. In this work, we assume that the FBSs are fully loaded as in [23], [24] because of the large number of users present in the overcrowded areas of their deployment. Offloading from the MBS tier adds more users to the FBSs. Therefore, the FBS may need to serve more than the commonly studied two-user in a NOMA group. Since, the FBS is assumed to be fully loaded, it is required to form a group of three (or more) users in order to serve the offloaded MU as shown in Fig. 1.

In this work, the numerical results suggest that the throughput of WU degrades with increase in number of users served in the NOMA group. This is because in NOMA signal of multiple users are superimposed. The weak users do not apply successive interference cancellation (SIC) on the message of the users with stronger channel gain. The messages of the users with stronger channel gain are treated as interference (called as intra-group interference). Hence, with increased users, the intra-group interference also increases, leading to degraded throughput of the WUs. Hence, in this work, we propose a novel time slotting (TS) technique that enhances performance of the WUs as the number of users increases in a NOMA group. Furthermore, the time slot duration and the power allocation coefficients are optimized to maximize the WU's sum rate.

The energy efficiency (EE) consideration for the 5G and beyond networks has turned out to be important concern since the information and communication technology accounts for nearly 5% of the entire world energy consumption [25]. Looking at the immense popularity gained by NOMA as an enabling technology for 5G and beyond, it is of interest to study the EE of the system while proposing any new methods in NOMA as analyzed in [26], [27]. In this work, the EE of the proposed TS techniques is investigated using the optimized time slot duration and the power allocation coefficients.

A. Difference from Existing Literature and Contributions

Major differences with existing literature and contribution of this work are:

- The authors in [16] achieve user fairness by improving performance of WUs through appropriate power allocation whereas [17] achieves performance enhancement of WU based on the selection of appropriate channel condition difference between the users paired in a NOMA group. In the proposed work, the authors introduce an additional design factor to guarantee fairness, called the time slot duration. Neither of the work in [16] and [17] discusses about performance of the WUs based on reducing the intra-group interference, which is integral to NOMA due to the superposition of the signal of multiple users. In this work, a novel TS technique is proposed which improves the WU's performance due to reduction in the intra-group interference (explained in Section III-B).
- To address the problem of increased intra-group interference in large NOMA groups, this work proposes a novel method using TS technique such that the number of users in a NOMA group is reduced by breaking the users into smaller NOMA groups. The smaller NOMA groups are served in different time slots, thereby lowering the intra-group interference at the WU in the NOMA group. Performance enhancement of WUs in the NOMA group is achieved by optimizing the time slot duration in which users are served, and by selecting the optimized power allocation coefficients for the NOMA group.
- The proposed TS techniques aims at user fairness by improving the throughput of WUs in a NOMA group while maintaining the minimum throughput of SU. In order to prove effectiveness of the proposed TS technique in terms of user fairness, QoS fairness index as well as the QoE fairness index are calculated as a measure for fairness amongst the users. The QoS fairness is measured using the commonly used Jain's fairness index while the QoE is measured using the MOS technique by considering a simple web page browsing scenario (explained in Section III-C3).
- The power allocation coefficients and the time slot duration are jointly optimized for the proposed TS technique. The numerical results thus obtained are compared with time division multiple access (TDMA), with the system model from [2] for a three-user NOMA group, and with the work presented in [17]. Furthermore, EE is calculated for the proposed TS techniques using the optimized power allocation coefficients and the time slot duration, and compared with the EE obtained by the conventional TDMA and with the EE achieved in [17].

B. Paper Organization

Rest of the paper is organized as follows: Section II discusses the proposed system model. Section II-A gives the expressions for the signal-to-interference-and-noise-ratio (SINR), and for the throughput at typical user (TU). Section III formulates the optimization problem and transforms the non-convex problem to a convex program, and introduces in detail the proposed TS techniques. Section IV analyses and discusses the numerical results obtained in detail. The paper concludes in Section V.

II. System Model

A two-tier network of MBSs and FBSs is considered which follow independent Poisson point process (PPP) based

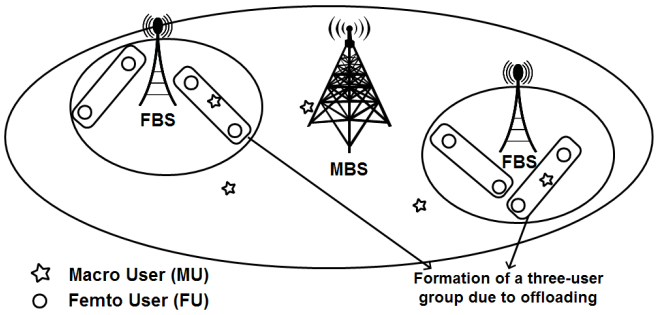

Fig. 1: System Model

distribution, Ω_t with density λ_t for the t^{th} tier such that $t \in \{m, f\}$, where m and f denote MBS tier and FBS tier, respectively. The FBS tier employs PD-NOMA to serve users. The transmit power of t^{th} tier is denoted by P_t. Bounded path loss model is considered as $L(r) = \frac{1}{(1+r_t^{\nu_t})}$, where ν_t is the path loss exponent for t^{th} tier and r_t represents the distance between the TU, and tagged BS that serves the TU of the t^{th} tier, respectively. Hence, the total channel gain for the TU is given by $|h_t|^2 = |\hat{h}_t|^2 L(r)$, where $\hat{h}_t$ follows Rayleigh fading. $\mathcal{R}_k$ is the target rate for k^{th} user and $\mathcal{Y}_t$ represents the communication range of BS of t^{th} tier. For tractable analysis, we consider a three-user NOMA group and divide the given time into two slots (explained in detail in Section III-B). It should be noted that the proposed TS can be extended to more than three users and more than two time slots. Schematic for the proposed TS is shown in Fig. 2.

A. SINR and Throughput at Typical Femto User

Initially, let us assume there are M users in a NOMA group and N time slots. The channel gains[2] of the M users of NOMA group are ordered as

$$|h_{1,n}|^2 \leq \ldots \leq |h_{M,n}|^2, \tag{1}$$

where $|h_{k,n}|^2$ denotes the total channel gain for k^{th} user in n^{th} time slot. We assume that the ordering remains same in all N-time slots. Given $x_{i,n}$ as the intended message for i^{th} user of the NOMA group such that $\mathbb{E}[x_{i,n}^2]$ are assumed to be equal, where $\mathbb{E}[\cdot]$ denotes the statistical expectation operator. The signal transmitted by the FBS in the n^{th} time slot is given by $X_{f,tx,n} = \sum_{i=1}^{M} x_{i,n}\sqrt{a_{i,n}P_f}$. Hence, the signal received by typical femto user (TFU), indexed k, of the NOMA group is given by $X_{f,rx,n} = h_{k,n}(\sum_{i=1}^{M} x_{i,n}\sqrt{a_{i,n}P_f}) + n_f$, where n_f is additive white Gaussian noise. User k decodes and removes message of all the users with channel gain weaker than itself. The message of the users with stronger channel gain is treated as interference while decoding its own message. The SINR at the TFU to decode the message of user j (such that $j < k$) for SIC is given as [3]

$$\gamma_{k,n}^{k\to j}(a_{j,n}) = \frac{\rho_{f,n} a_{j,n} |h_{k,n}|^2}{\rho_{f,n}|h_{k,n}|^2 \sum_{l=j+1}^{M} a_{l,n} + \sum_t \rho_t^I \mathcal{I}_{t,n} + 1}, \tag{2}$$

where $\rho_{f,n} = \mathbb{E}[x_{i,n}^2]/\sigma_f^2$ denotes the transmit signal-to-noise-ratio (SNR) at FBS in the n^{th} time slot, σ_f^2 denotes the noise variance, $a_{u,n}$ denotes power allocation coefficients for user with index $u = \{k, j, l\}$ in the n^{th} time slot, $\rho_t^I = P_t/\sigma_f^2$ denotes the transmit SNR from the transmitter of t^{th} tier responsible for interference, and $\rho_t^I \mathcal{I}_{t,n}$ denotes the interference from t^{th} tier in the n^{th} time slot. Assuming the TFU at the origin according to the Slivnyak's theorem [28] and the tagged FBS at f_0, $\mathcal{I}_{f,n} = \rho_f^I \sum_{i\in\Omega_f/\{f_0\}} |\tilde{h}_{i,n}|^2$ is the co-tier interference at the TFU, where $|\tilde{h}_{i,n}|^2$ denotes the total channel gain from i^{th} FBS to the TFU in the n^{th} time slot, and $\mathcal{I}_m = \rho_m^I \sum_{i\in\Omega_m} |\tilde{h}_{i,n}|^2$ is the cross-tier interference at the TFU in the n^{th} time slot from the MBS tier, where $|\tilde{h}_{j,n}|^2$ represents the total channel gain from j^{th} MBS to the TFU in the n^{th} time slot. The corresponding throughput required at the TFU in the n^{th} time slot to successfully decode message of user j ($j < k$) can be calculated as

$$\mathcal{R}_{k,n}^{k\to j}(a_{j,n}) = \log(1 + \gamma_{k,n}^{k\to j}(a_{j,n})). \tag{3}$$

The SINR at the TFU of the NOMA group to decode its own message in the n^{th} time slot is given by

$$\gamma_{k,n}(a_k) = \frac{\rho_{f,n} a_{k,n} |h_{k,n}|^2}{\rho_{f,n}|h_{k,n}|^2 \sum_{l=k+1}^{M} a_{l,n} + \sum_t \rho_t^I \mathcal{I}_{t,n} + 1}. \tag{4}$$

The corresponding throughput required at the TFU to decode its own message in the n^{th} time slot is calculated as

$$\mathcal{R}_{k,n}(a_{k,n}) = \log(1 + \gamma_{k,n}(a_{k,n})). \tag{5}$$

III. Optimization of Power Allocation Coefficients and Time Slot Duration

This work aims to enhance the throughput of WUs in a NOMA group. In order to achieve this objective, the sum throughput of the WUs in a NOMA group is maximized by optimizing the power allocation coefficients and time slot duration for the proposed TS. Hence, we are interested in jointly optimizing time slot duration (t_n), and the power allocation coefficients ($a_{k,n}$) to maximize the sum throughput of WUs in a NOMA group. It should be noted that the optimized values are calculated individually for different techniques of the proposed TS. Assuming $\mathcal{K}$ denotes the set of WUs in a NOMA group, the optimization problem aims at maximizing the sum throughput for the set $\mathcal{M} = \mathcal{K} \times \{1, 2, ..., N\}$ (explained in detail in Section III-B). The optimization problem can be

[2]Throughout the paper, $\hat{h}$ denotes Rayleigh distribution, $|\tilde{h}|^2$ denotes the unordered channel gain, and $|h|^2$ denotes ordered channel gain.

TABLE I: Notations and their values used in the numerical analysis

Parameter	Description	Value
α_n	Additional variables used to make the objective function convex	-
$\gamma_{k,n}^{k\to j}(a_{j,n})$	SINR at k^{th} user to decode the message of j^{th} user in n^{th} time slot	-
$\gamma_{k,n}(a_k)$	SINR at k^{th} user to decode its own message	-
η	Energy Efficiency	-
λ_m and λ_f	MBS tier and FBS tier density, respectively	5×10^{-5}, 1×10^{-4}
ν_m and ν_f	Path loss exponent for MBS and FBS tier, respectively	$3, 4$
ρ_f, ρ_t^I	Transmit SNR at FBS, Interfering SNR from t^{th} tier	-
σ_f^2	Noise variance	1
Ω_t	PPP distribution for t^{th} tier	-
M	Number of users	3
N	Number of time slots	2
P_f	Transmitting power for FBS tier	1 W
R_k	Target data rate of k^{th} user	0.1 bps
$X_{f,tx,n}$	Superimposed signal transmitted by FBS	-
$X_{f,rx,n}$	Superimposed signal received by the user	-
$a_{k,n}$	Power allocation coefficient for k^{th} user in n^{th} time slot	Optimized
$\hat{h}$, $\lvert\tilde{h}\rvert^2$	Rayleigh distributed channel gain, Unordered total channel gain	-
$\lvert h_{i,n}\rvert^2$	Ordered total channel gain for i^{th} user in n^{th} time slot	-
n_f	Additive white Gaussian noise	-
t_n	Time duration of the n^{th} time slot	Optimized
$x_{k,n}$	Intended signal for k^{th} user in n^{th} time slot	-
$\mathbb{E}[\cdot]$	Statistical expectation operator	-
$\mathcal{E}_{web}$ and $\mathcal{J}$	QoE and Jain's (QoS) fairness index	-
$\mathcal{G}_{k,n}$, $\mathcal{G}_{k,n}^{(f)}(a_{k,n},\alpha_n)$	Non-convex objective function, Convex objective function	-
$\mathcal{I}_{t,n}$	Interference from the t^{th} tier in n^{th} time slot	-
$\mathcal{K}$	Set consisting of WU's index	-
$\mathcal{M}$	Set consisting of WU's index in n^{th} time slot	-
$\mathcal{R}_{k,n}(a_{k,n})$	Throughput at the k^{th} user in n^{th} time slot	-
$\mathcal{Y}_m$ and $\mathcal{Y}_f$	Transmission range for MBS and FBS tier, respectively	1000m, 5m

formulated as

$$\max_{a_{k,n},t_n} \sum_{k,n\in\mathcal{M}} t_n\times\mathcal{R}_{k,n}(a_{k,n}) \tag{6a}$$

$$\text{s.t.}\quad t_n\times\mathcal{R}_{k,n}(a_{k,n})\geq R_k, \tag{6b}$$

$$\sum_{k=1}^{M} a_{k,n}\leq P_f, a_{k,n}\geq 0, \text{and} \tag{6c}$$

$$\sum_{n=1}^{N} t_n\leq T, t_n\geq 0, \tag{6d}$$

where T denotes the total time duration. To begin, additional variables α_n are introduced [29], [30] which satisfy the following convex constraints

$$\alpha_n t_n\geq 1 \text{ and } \sum_{n=1}^{N} 1/\alpha_n\leq 1. \tag{7}$$

The problem (6) can equivalently be written as

$$\max_{a_{k,n},t_n} \mathcal{G}_{k,n}=\sum_{k,n\in\mathcal{M}} t_n\times\mathcal{R}_{k,n}(a_{k,n}) \tag{8a}$$

$$\text{s.t.}\quad t_n\times\mathcal{R}_{k,n}(a_{k,n})\geq \mathcal{R}_k, \tag{8b}$$

$$\sum_{k=1}^{M} a_{k,n}\leq P_f, a_{k,n}\geq 0, \text{and} \tag{8c}$$

$$\alpha_n t_n\geq 1 \text{ and } \sum_{n=1}^{N} 1/\alpha_n\leq 1. \tag{8d}$$

The objective function in (8a) is non-concave. Also, the constraint in (8b) is a non-convex constraint. Considering first the non-concave objective function, (4) can be equivalently written as

$$\gamma_{k,n}(a_{k,n})=\frac{\left(h_{k,n}\sqrt{\rho_{f,n}a_{k,n}}\right)^2}{\zeta_{k,n}(a_{k,n})}, \tag{9}$$

where $\zeta_{k,n}(a_{k,n}) = \rho_{f,n}|h_{k,n}|^2 \sum_{l=k+1}^{M} a_{l,n} + \sum_t \rho_t^I \mathcal{I}_{t,n} + 1$. This gives an additional linear constraint as follows

$$h_{k,n}\sqrt{\rho_{f,n} a_{k,n}} \geq 0. \tag{10}$$

Furthermore, the following inequalities are used similar to [29], [30], the proof of which are given in Appendix A. For a given point (x^f,y^f), we may approximate $\log\left(1+\frac{|x|^2}{y}\right)$ as

$$\log\left(1+\frac{|x|^2}{y}\right) \geq \log\left(1+\frac{|x^f|^2}{y^f}\right) - \frac{|x^f|^2}{y^f} + 2\frac{(x^f)^* x}{y} - \frac{|x^f|^2(|x|^2+y)}{y^f(y^f+|x^f|^2)}, \tag{11}$$

$$\frac{|x|^2}{y} \geq 2\frac{(x^f)^* x}{y^f} - \frac{|x^f|^2}{{y^f}^2} y, \tag{12}$$

$$\forall x \in \mathbb{C}, x^f \in \mathbb{C}, y > 0, y^f > 0.$$

where, $(\cdot)^*$ denotes the complex conjugate operator. Hence, at a feasible point $(a_{k,n}^{(f)}, t_n^{(f)})$, following from the inequality (11) we write the non-concave objective function in (8(a)) as

$$\log\left(1+\gamma_{k,n}(a_{k,n})\right) \geq \log\left(1+\gamma_{k,n}(a_{k,n}^{(f)})\right) - \gamma_{k,n}\left(a_{k,n}^{(f)}\right) + 2\frac{\omega_{k,n}^{(f)}\ \omega_{k,n}}{\zeta_{k,n}(a_{k,n}^{(f)})} - \frac{\left(\omega_{k,n}^{(f)}\right)^2\left(\zeta_{k,n}(a_{k,n}) + (\omega_{k,n})^2\right)}{\zeta_{k,n}(a_{k,n}^{(f)})\left(\zeta_{k,n}(a_{k,n}^{(f)}) + \left(\omega_{k,n}^{(f)}\right)^2\right)}, \tag{13}$$

where, $\omega_{k,n}^{(f)} = h_{k,n}\sqrt{\rho_f a_k^{(f)}}$. Setting the values as $x_{k,n}^{(f)} = \log\left(1+\gamma_{k,n}(a_{k,n}^{(f)})\right) - \gamma_{k,n}\left(a_{k,n}^{(f)}\right)$, $y_{k,n}^{(f)} = 2\frac{\omega_{k,n}^{(f)}}{\zeta_{k,n}(a_{k,n}^{(f)})}$, and $z_{k,n} = \frac{\left(\omega_{k,n}^{(f)}\right)^2}{\zeta_{k,n}(a_{k,n}^{(f)})\left(\zeta_{k,n}(a_{k,n}^{(f)})+\left(\omega_{k,n}^{(f)}\right)^2\right)}$, from (12) and (13) we get

$$\begin{aligned}
&\frac{\log\left(1+\gamma_{k,n}(a_k)\right)}{\alpha_n} \\
&\geq \frac{x_{k,n}^{(f)}}{\alpha_n} + y_{k,n}^{(f)}\frac{\omega_{k,n}^{(f)}}{\alpha_n} - z_{k,n}\frac{\left(\zeta_{k,n}(a_{k,n}) + (\omega_{k,n})^2\right)}{\alpha_n} \\
&\geq \frac{x_{k,n}^{(f)}}{\alpha_n} + y_{k,n}^{(f)}\left(2\frac{\sqrt{\omega_{k,n}^{(f)}}\sqrt{\omega_{k,n}}}{\alpha_n^{(f)}} - \frac{\omega_{k,n}^{(f)}\alpha_n}{(\alpha_n^{(f)})^2}\right) \\
&- z_{k,n}\frac{\left(\zeta_{k,n}(a_{k,n}) + (\omega_{k,n})^2\right)}{\alpha_n} = \mathcal{G}_{k,n}^{(f)}(a_{k,n},\alpha_n).
\end{aligned} \tag{14}$$

The function $\mathcal{G}_{k,n}^{(f)}(a_{k,n},\alpha_n)$ in (14) is concave and is the global lower bound of $\frac{\log(1+\gamma_{k,n}(a_{k,n}))}{\alpha_n}$. Initialized by feasible point $(a_{k,n}^{(0)}, \alpha_n^{(0)})$ for optimization problem in (8), the convex optimization problem in (15) is solved at the f^{th} iteration to generate the next feasible point $a_{k,n}^{(f+1)}, \alpha_n^{(f+1)}$. The procedure for finding the initial point $(a_{k,n}^{(0)}, \alpha_n^{(0)})$ for optimization problem is discussed in Section III-A.

$$\max_{a_{k,n},\alpha_n} \sum_{k,n\in\mathcal{M}} \mathcal{G}_{k,n}^{(f)}(a_{k,n},\alpha_n) \tag{15a}$$

$$\text{s.t.} \quad \mathcal{G}_{k,n}^{(f)}(a_{k,n},\alpha_n) \geq R_k, \tag{15b}$$

$$\sum_{k=1}^{M} a_{k,n} \leq P_f, a_{k,n} \geq 0, \tag{15c}$$

$$\alpha_n t_n \geq 1 \text{ and } \sum_{n=1}^{N} 1/\alpha_n \leq 1, \text{and} \tag{15d}$$

$$|h_{k,n}|\sqrt{\rho_{f,n} P_f a_{k,n}} \geq 0. \tag{15e}$$

Proposition 1: The sequence $\{(a_{k,n}^{(f+1)}, \alpha_n^{(f+1)})\}$ of improved feasible points for (8) thus converges at least to a locally optimal solution satisfying the Karush-Kuhn-Tucker conditions [31].

Proof: Please see Appendix B.

A. Generating initial feasible point

To get the initial point for the optimization problem in (15), we start from a feasible point $(a_{k,n}^{(0)}, \alpha_n^{(0)})$ for constraint (7), and iterate the following convex program

$$\max_{a_{k,n},\alpha_n} \min_{(k,n)\in\mathcal{M}} \mathcal{G}_{k,n}^{(f)}(a_{k,n},\alpha_n)/\mathcal{R}_k \tag{16a}$$

$$\text{s.t.} \quad \alpha_n t_n \geq 1 \text{ and } \sum_{n=1}^{N} 1/\alpha_n \leq 1, \text{ and} \tag{16b}$$

$$h_{k,n}\sqrt{\rho_{f,n} P_f a_{k,n}} \geq 0. \tag{16c}$$

The convex program in (16) is iterated till $\min_{(k,n)\in\mathcal{M}} \mathcal{G}_{k,n}^{(f)}(a_{k,n}^{f+1}, \alpha_n^{f+1})/\mathcal{R}_k \geq 1$ is reached. When the condition $\min_{(k,n)\in\mathcal{M}} \mathcal{G}_{k,n}^{(f)}(a_{k,n}^{f+1}, \alpha_n^{f+1})/\mathcal{R}_k \geq 1$ is met, the program is terminated and the point $(a_{k,n}^{f+1}, \alpha_n^{f+1})$ is obtained. The imposed condition $\min_{(k,n)\in\mathcal{M}} \mathcal{G}_{k,n}^{(f)}(a_{k,n}^{f+1}, \alpha_n^{f+1})/\mathcal{R}_k \geq 1$ makes $(a_{k,n}^{f+1}, \alpha_n^{f+1})$ feasible for (8) and thus is used as an initial feasible point for (15).

From the discussion given in [29], [30], the optimization problem in (15) involves $a = 2(3M+1)$ quadratic and linear constraints, and $b = 2(M+1)$ decision variables, hence, the computational complexity is given as $\mathcal{O}(a^2 b^{2.5} + b^{3.5})$.

B. Proposed TS Technique

In this section, we discuss in detail how the users are served using the proposed TS technique by breaking a large NOMA group into smaller NOMA groups such that each smaller NOMA group is served in different time slots. In order to clearly demonstrate the gains achieved by using the proposed approach, without loss of generality, this work consider a limiting use case of three NOMA users and serve them in two time slots. We assume that the FBS needs to serve a three-user NOMA group due to the offloading from the MBS tier. Let the channel gain of the three-user NOMA group in n^{th} time slot be ordered as $|h_{1,n}|^2 < |h_{2,n}|^2 < |h_{3,n}|^2$. Further, based

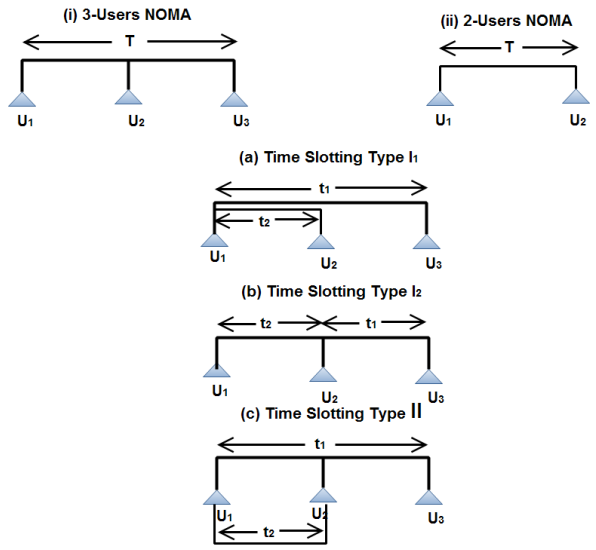

Fig. 2: Proposed TS techniques in a NOMA group.

on the order of the channel gain of the users, in this work, U_3 is assumed as the SU, and U_2 and U_1 as the WUs in the NOMA group. The three TS techniques are as follows:

- TS Type I_1: As shown in Fig. 2 (a), in the first time slot t_1, U_3 is paired with U_1, and is served as a two-user NOMA group. Later, in the next time slot t_2, U_1 and U_2 are paired and are served as two-user NOMA group. Therefore, assuming that U_1 is a WU, it is served for the whole time in two different two-user NOMA groups rather than serving for the whole time in a three-user NOMA group. The optimization problem aims at maximizing the throughput of U_1 such that $\mathcal{M} = \{1\} \times \{1, 2\}$. The set $\mathcal{M}$ denotes the WU targeted in TS Type I_1 is U_1 and the sum rate is calculated over the two time slots.
- TS Type I_2: As shown in Fig. 2 (b), in the first time slot t_1, U_2 and U_3 are served while in the next slot t_2, U_1 and U_2 are served using two-user NOMA group. Since U_2 is a WU, it is alloted two time slots, similar to TS Type I_1. Therefore, the optimization problem aims at maximizing the throughput of U_2 such that $\mathcal{M} = \{2\} \times \{1, 2\}$. Similar to TS Type I_1, here the set $\mathcal{M}$ indicates that the WU targeted in TS Type I_2 is U_2 and the sum rate is calculated over the two time slots.
- TS Type II: As shown in Fig. 2 (c), in the first time slot t_1, all the three users are served using a three-user NOMA group, while in the next time slot t_2, a two-user NOMA group is used to serve user U_1 and U_2. This implies that both the WUs, U_1 and U_2, are served in two time slots. Therefore, the optimization problem aims at maximizing the sum throughput of U_2 and U_1 such that $\mathcal{M} = \{1, 2\} \times \{1, 2\}$. Here, unlike TS Type I_1 and TS Type I_2, both the WUs are considered as users with high priority data, as denoted by the set $\mathcal{M}$.

Note: It should be noted that, the method of grouping/scheduling of users in NOMA groups is left for future study. This work studies in detail the TS methods. For a three-user and two time slots case, the above three TS technique covers all the possible combinations. The same can be extended for higher number of users and time slots. Also, for future work, combinational optimization technique can be explored which lists all the combinations and chooses the best one according to the requirement. Such combinational optimization technique will also be useful when the number of users and/or number of time slots increases.

C. EE, QoS Fairness Index and QoE Fairness Index Calculation

This section calculates the EE of the system using the proposed TS techniques. Furthermore, to show the user fairness achieved by the proposed TS technique, two fairness index, one to measure the QoS fairness and one to measure the QoE fairness are calculated and discussed in detail. The QoS fairness is measured using the Jain's fairness index while the MOS technique is considered to calculate the QoE fairness by considering a simple scenario of web page browsing.

1) EE Calculation: EE is calculated as the ratio of sum rate achieved by user over the total power consumed [32], [33]. The EE achieved by the different TS technique is calculated as

$$\eta = \frac{\sum\limits_{k,n \in \mathcal{M}} \mathcal{G}_{k,n}}{P_f + P_c} \tag{17}$$

where, P_c denotes the additional power consumed in the circuit.

2) QoS Fairness Index: In the existing literature one of the commonly used QoS fairness index for wireless networks is Jain's fairness index [7], [8]. Hence, in this work, Jain's fairness index is used as a measurement of the QoS fairness. The Jain's fairness index for the proposed TS techniques can be expressed as

$$\mathcal{J} = \frac{\left(\sum\limits_{k,n \in \mathcal{M}} \mathcal{G}_{k,n} \right)^2}{\left(M \times \sum\limits_{k,n \in \mathcal{M}} \mathcal{G}_{k,n}^2 \right)}. \tag{18}$$

3) QoE Fairness Index using MOS Model for Web Browsing: The QoE is an important criterion to assure that the users are satisfied by the received service. The MOS model is used in the literature to predict the QoE experienced by the users. Since, web browsing is the most commonly used application in the wireless networks, in this work, the users' QoE is calculated based on its experience while browsing a web page. The MOS model for web browsing application is defined similar to [11] as follows:

$$\mathcal{E}_{web} = -C_1 \ln(d(\mathcal{R})) + C_2, \tag{19}$$

where $\mathcal{R}$ is the rate achieved by the user. $\mathcal{E}_{web}$ denotes the score which ranges between 1 to 5. This score reflects the quality experienced by the user. A score of 5 implies best quality at the user while score of 1 denotes that the quality experienced by the user is worst. The constants C_1

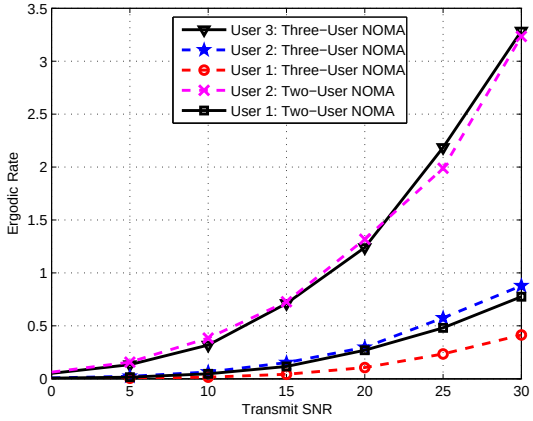

Fig. 3: Throughput for $M = 2$ and $M = 3$.

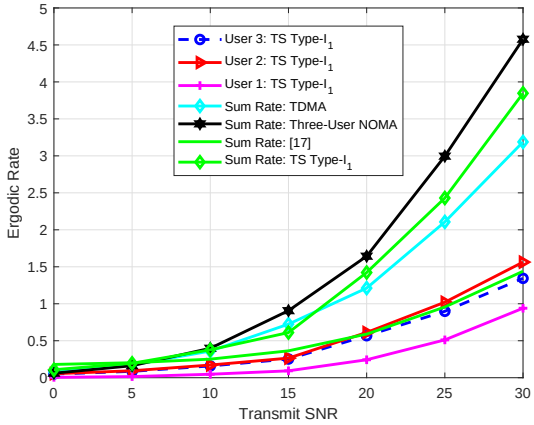

Fig. 4: Throughput for TS Type I_1.

and C_2 are decided based on the experiments on the web browsing applications and are set to be 1.120 and 4.6746, respectively [11]. $d(\mathcal{R})$ is the delay time. The delay time denotes the time taken between sending of request by the user for a web page and displaying of the web page contents. Based on the assumption in [11], delay time can simplified as $d(\mathcal{R}) = \mathcal{R}/FS$, where FS is the frame size.

IV. Results and Discussions

This section evaluates throughput for the proposed TS techniques. The MATLAB and Statistics Toolbox [34] are used for solving the optimization problem. The optimized values are derived by averaging the simulation results over 10^5 iterations using PPP distribution of BSs in the disc with an indicative disc radius of 1000m. The parameters are taken to be $R_i = 0.1$ bps $\forall\{i = 1, 2, 3\}$, $\lambda_m = 5 \times 10^{-5}$, $\lambda_f = 10^{-4}$, $P_m = 20$W, $P_f = 1$W, $\mathcal{Y}_m = 1000$m, $\mathcal{Y}_f = 5$m, $\nu_m = 3$, $\nu_f = 4$, and $FS = 800$ kB [11]. Comparison of the individual throughputs of the users achieved by the proposed TS technique is done with that achieved using three-user NOMA from the system model in [2]. Comparison of the sum throughput achieved by the proposed TS technique is done between the conventional TDMA, three-user NOMA user and with the work presented in [17]. Furthermore, the EE of the proposed TS technique is compared with the conventional TDMA technique and with the EE obtained in [17].

We assume that in a three-user NOMA group the channel gain of the users are ordered as follows in the n^{th} time slot and the order remains the same in all the time slots.

$$|h_{n,1}|^2 < |h_{n,2}|^2 < |h_{n,3}|^2. \tag{20}$$

Further to distinguish the users we consider the user with the best channel gain, i.e., U_3 as the SU and the remaining two users as the WUs. Furthermore, U_1 is considered as the weakest user since it has the worst channel gain amongst the three users.

Note: The objective of this work is to enhance performance of the WUs when the number of users in a NOMA group increases and achieve user fairness. Therefore, this work targets comparison of WU's performance between three-user NOMA and the proposed TS technique. This work do not aim to achieve better performance than OMA, since, due to increased number of users, it may not be possible to dedicate an entire band to a user due to resource constraints.

As can be observed that while comparing the individual throughputs for three-user NOMA group and two-user NOMA group, it is evident from Fig. 3 that the SU achieves almost the same throughput whether using three-user NOMA or using two-user NOMA. On the other hand, while comparing the throughput of WUs, user-1 of a three-user NOMA group experiences degradation of 46.56% in its throughput as compared to if it was served as a WU using two-user NOMA group. This proves that in terms of the throughput achieved by the WU, two-user NOMA is preferable as compared to three-user NOMA. The reason for lower throughput of WUs in a three-user NOMA as compared to two-user NOMA is that as the number of users increases in a NOMA group, the intra-group interference increases due to the superposition of signal of multiple users in NOMA. The weaker users do not use SIC to remove the signal of the users with stronger channel gain but, treat it as interference while decoding their own signal. The increased intra-group interference results in degraded throughput at the WUs. Hence, in this work, a novel TS technique is proposed that improves performance of the WUs, in case the FBS tier is required to serve more than two users. The need to serve more than two users arise when an MU is offloaded to the FBS tier due to congestion at the MBS tier.

Fig. 4 shows the first TS technique, TS Type I_1. In TS Type I_1, instead of serving the three users using a three-user NOMA, the users are served using a two-user NOMA in two separate time slots as shown in Fig. 2(a). Breaking the three-user NOMA into two time slots reduces the degradation in throughput at the WU caused due to the increased intra-group interference from the three-user NOMA. Contrary to TDMA which can serve only two users in two time slots, using the proposed TS in NOMA, three users are served in two time slots. Using TS Type I_2, U_2 achieves throughput enhancement by 77.90%, while U_1 achieves a significant performance enhancement in throughput by 126.90%. Also, using three-

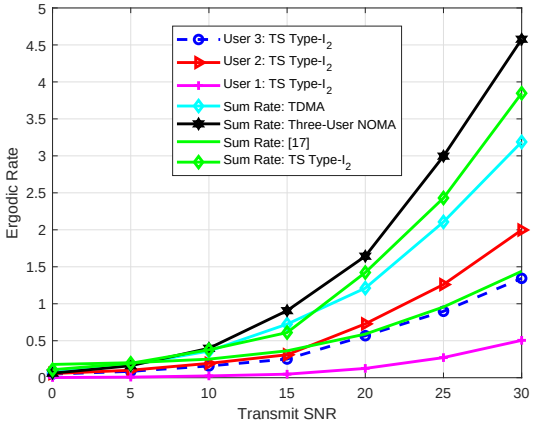

Fig. 5: Throughput for TS Type I_2.

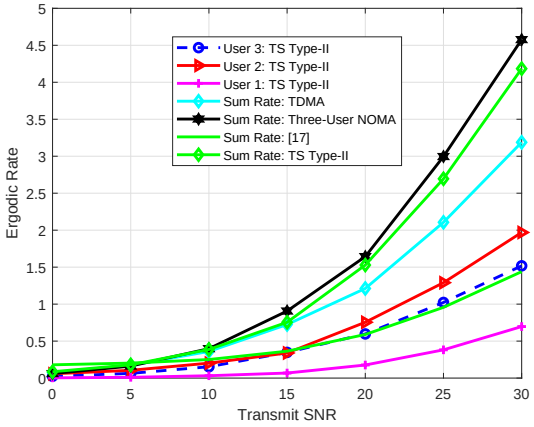

Fig. 6: Throughput for TS Type II.

user NOMA the difference in the throughput between the strongest user, i.e., U_3 and U_2 is 73.25%, and between U_3 and the weakest user, i.e., U_1 is 87.39%. Using the TS Type I_1, the difference in throughput between U_3 and U_2 reduces to 16.29%, while between U_3 and U_1 reduces to 30.09%. Hence, using TS Type I_1 proves to enhance the throughput of the WUs and also achieves user fairness. Detailed discussion on the QoS and QoE fairness achieved by the proposed TS technique is presented in Fig. 7 and Fig. 8, respectively. Moreover, the sum throughput is enhanced by 20.67% by using the TS Type I_1 as compared to the sum throughput when conventional TDMA is used.

Fig. 5 shows the second TS technique, TS Type I_2 which is nearly similar to TS Type I_2, however instead of the U_1, two time slots are dedicated to U_2, since, U_2 is assumed to have high priority data. TS Type I_2 enhances the throughput of U_1 by 22.05% and improves the throughput of U_2 by 127.41%. Also, the difference in performance between U_3 and U_2 reduces to 48.66% and between U_3 and U_1 reduces to 62.40% as compared with three-user NOMA. The sum throughput by using TS Type I_2 improves by 20.67% as compared to using TDMA. As observed from Fig. 4 and Fig. 5, the sum throughput of TS Type I_1 and TS Type I_2 achieve improvement of 167.26% as compared to that achieved in [17].

Fig. 6 shows the throughput by using the third TS technique, TS Type II,wherein we use a combination of three-user NOMA and two-user NOMA in the two time slots. Using TS Type II, U_2 achieves throughput enhancement of 124.11% and U_1 achieves an improvement of 68.29% in the throughput. Moreover, using TS Type II reduces performance difference between the SU and the WU leading to a difference of 29.71% between U_3 and U_2. The difference between performance of U_3 and U_1 by using TS Type II is reduced to 54.10% when compared to three-user NOMA. Furthermore, TS Type II, achieves highest sum throughput and achieves an improvement of 31.24% as compared to the conventional TDMA scheme. The sum throughput of TS Type II achieves improvement of 211.87% in comparison to the sum throughput attained in [17]. Also, the sum throughput with TS Type II is nearly the same as that of the three-user NOMA.

Fig. 7 shows the QoS fairness based on Jain's fairness index and Fig. 8 depicts the QoE fairness based on the MOS model for web browsing application. The QoS and the QoE fairness for the proposed TS technique is compared with the conventional three-user NOMA. The numerical values of the Jain's fairness index from Fig. 7 demonstrates that all the proposed TS techniques achieves significantly better QoS fairness as compared to using three-user NOMA. The QoS fairness between the users improves by 61.43% while using TS Type I_1 and by 36.92% when TS Type I_2 is used over three-user NOMA. Similarly, an improvement of 47% is observed when TS Type II is used as compared to using three-user NOMA. Fig. 8 depicts the QoE of the proposed TS technique by considering the common application of web browsing. It can be observed for Fig. 8 that the QoE fairness achieved by the all the proposed TS techniques is better than that achieved by three-user NOMA. The QoS fairness between the users improves by 5.07% while using TS Type I_1 and by 3% when TS Type I_2 is used over three-user NOMA. Similarly, an improvement of 4% is observed when TS Type II is used as compared to using three-user NOMA. Improvement in both QoS fairness and QoE fairness justifies the use of the proposed work for achieving fairness amongst the users when the number of users increases to be served using NOMA.

Fig. 9 shows the EE of the conventional TDMA and proposed TS technique using the optimized values calculated in Section III. The results demonstrate that the EE achieved by the proposed TS techniques is higher as compared to conventional TDMA. The TS Type I_1 and TS Type I_2 achieves improved EE by 30.71% as compared to TDMA. The EE achieved by TS Type II shows an improvement of 52.57% as compared to TDMA. It can also be inferred from Fig. 9 that the EE of TS Type I_1 and TS Type I_2 improve by 78.23% as compared to that achieved in [17]. The EE of TS Type II improves by 107.99% in comparison to the EE attained in [17].

Hence, the numerical results prove that using the proposed TS technique aids in performance enhancement of the WU and achieves higher QoS and QoE fairness between users. Furthermore, the TS techniques improves the EE of the system over the conventional TDMA.

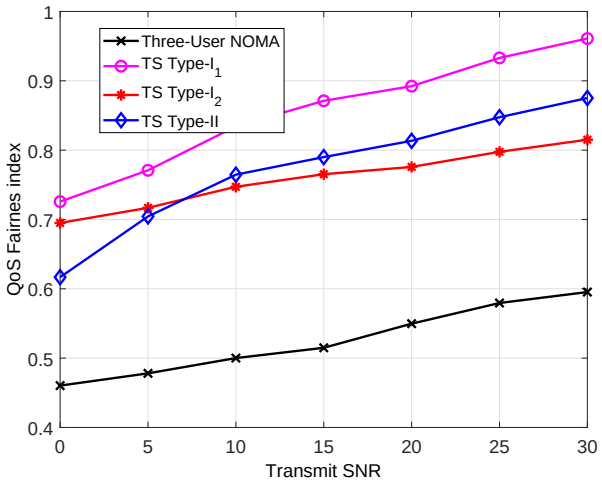

Fig. 7: QoS Fairness index

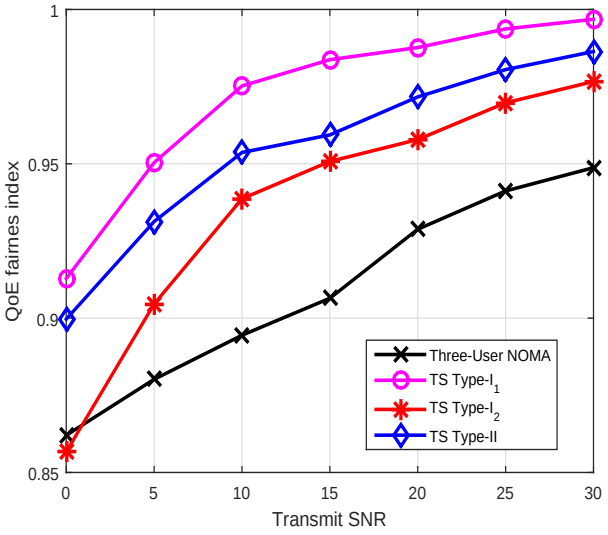

Fig. 8: QoE Fairness Index.

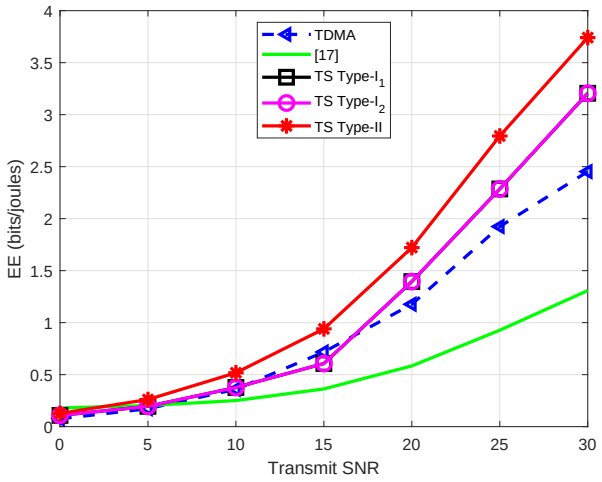

Fig. 9: EE achieved by TDMA and proposed TS techniques.

V. Conclusion

The paper proposes novel TS techniques for a three-user NOMA group to enhance throughput of the WU. An optimization technique is proposed to jointly optimize the power allocation coefficients and time slot duration for the proposed TS techniques. The proposed TS techniques doubles the throughput of WUs in certain cases due to decreased intra-group interference while maintaining the minimum throughput requirement of the SU. Additionally, the three proposed TS techniques achieve better user fairness in comparison to three-user NOMA. Furthermore, the proposed TS technique also attains better EE as compared to the conventional TDMA.

Appendix A
Proof of Inequality (11) and (12)

The function $f(t) = -\log(1-t)$ is convex and increasing in the domain $0 \leq t < 1$, while the function $g(x,z) = \frac{|x|^2}{z}$ is convex. Therefore, the composite function given as

$$f(g(x,z)) = -\log(1-\frac{|x|^2}{z}), \tag{21}$$

is convex in the domain $z > |x|^2$ [35]. For a given point x^f and z^f, the following relation holds as given in [35]

$$-\ln\left(1-\frac{|x|^2}{z}\right) \geq -\ln\left(1-\frac{|x^f|^2}{z^f}\right) \tag{22}$$

$$-\frac{|x^f|^2}{z^f-|x^f|^2} + 2\frac{(x^f)^* x}{z^f-|x^f|^2} \tag{23}$$

$$-\frac{|\bar{x}|^2 z}{(\bar{z}-|\bar{x}|^2)\bar{z}}. \tag{24}$$

By noting the following relation

$$\ln\left(1+\frac{|x|^2}{y}\right) = -\ln\left(1-\frac{|x|^2}{y+|x|^2}\right) \tag{25}$$

(11) is obtained by applying (25) for $z = y + |x|^2$ and $z^f = y^f + |x^f|^2$. Furthermore, as $g(x,y) = \frac{|x|^2}{y}$ is convex in x and $y > 0$, it is true from [35] that,

$$\frac{|x|^2}{y} \geq 2\frac{(x^f)^* x}{y^f} - \frac{|x^f|^2}{y^{f^2}} y \tag{26}$$

Equation (12) is obtained by using (26). This completes the proof.

Appendix B
Proof of Proposition 1

Note that $\mathcal{G}_{k,n}(a_{k,n},\alpha_n) \geq \mathcal{G}^{(f)}_{k,n}(a_{k,n},\alpha_n)\ \forall\ (a_{k,n},\alpha_n)$, and $\mathcal{G}_{k,n}(a^{(f)}_{k,n},\alpha^{(f)}_n) = \mathcal{G}^{(f)}_{k,n}(a^{(f)}_{k,n},\alpha^{(f)}_n)$. Moreover, $\mathcal{G}^{(f)}_{k,n}(a^{(f+1)}_{k,n},\alpha^{(f+1)}_n) > \mathcal{G}^{(f)}_{k,n}(a^{(f)}_{k,n},\alpha^{(f)}_n)$ whenever $(a^{(f+1)}_{k,n},\alpha^{(f+1)}_n) \neq (a^{(f)}_{k,n},\alpha^{(f)}_n)$ because the former and the latter, respectively, are the optimal solution and feasible point for (15). Therefore, $\mathcal{G}_{k,n}(a^{(f+1)}_{k,n},\alpha^{(f+1)}_n) \geq \mathcal{G}^{(f)}_{k,n}(a^{(f+1)}_{k,n},\alpha^{(f+1)}_n) > \mathcal{G}^{(f)}_{k,n}(a^{(f)}_{k,n},\alpha^{(f)}_n) = \mathcal{G}_{k,n}(a^{(f)}_{k,n},\alpha^{(f)}_n)$, showing that $(a^{(f+1)}_{k,n},\alpha^{(f+1)}_n)$ is a better feasible point than $(a^{(f)}_{k,n},\alpha^{(f)}_n)$ for (8).

This completes the proof that the sequence $\{(a_{k,n}^{(f+1)}, \alpha_n^{(f+1)})\}$ of improved feasible points for (8) thus converges at least to a locally optimal solution satisfying the Karush-Kuhn-Tucker conditions.

REFERENCES

[1] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *IEEE Veh. Technol. Conf. (VTC Spring), 2013*, pp. 1–5.

[2] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, 2014.

[3] Y. Liu, Z. Ding, M. Elkashlan, and J. Yuan, "Non-orthogonal multiple access in large-scale underlay cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 10 152–10 157, 2016.

[4] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, 2016.

[5] H. Xing, Y. Liu, A. Nallanathan, Z. Ding, and H. V. Poor, "Optimal throughput fairness tradeoffs for downlink non-orthogonal multiple access over fading channels," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3556–3571, 2018.

[6] ——, "Optimal throughput fairness tradeoffs for downlink non-orthogonal multiple access over fading channels," *IEEE Trans. Wireless Commun.*, vol. 17, no. 6, pp. 3556–3571, 2018.

[7] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination," *Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA*, 1984.

[8] L. Chen, L. Ma, and Y. Xu, "Proportional fairness-based user pairing and power allocation algorithm for non-orthogonal multiple access system," *IEEE Access*, vol. 7, pp. 19 602–19 615, 2019.

[9] S. Huaizhou, R. V. Prasad, E. Onur, and I. Niemegeers, "Fairness in wireless networks: Issues, measures and challenges," *IEEE Commun. Surveys & Tut.*, vol. 16, no. 1, pp. 5–24, 2013.

[10] T. Hoßfeld, L. Skorin-Kapov, P. E. Heegaard, and M. Varela, "Definition of QoE fairness in shared systems," *IEEE Commun. Lett.*, vol. 21, no. 1, pp. 184–187, 2016.

[11] J. Cui, Y. Liu, Z. Ding, P. Fan, and A. Nallanathan, "QoE-based resource allocation for multi-cell NOMA networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 6160–6176, 2018.

[12] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, 2015.

[13] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, 2016.

[14] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, 2017.

[15] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, 2016.

[16] S. Timotheou and I. Krikidis, "Fairness for non-orthogonal multiple access in 5G systems," *IEEE Signal Process. Lett.*, vol. 22, no. 10, pp. 1647–1651, 2015.

[17] P. Swami, V. Bhatia, S. Vuppala, and T. Ratnarajah, "User fairness and performance enhancement for cell edge user in NOMA-HCN with offloading," in *IEEE Vehicular Technology Conference (VTC Spring)*, 2017, pp. 1–5.

[18] ——, "A cooperation scheme for user fairness and performance enhancement in NOMA-HCN," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 11 965–11 978, 2018.

[19] X. Gao, P. Wang, D. Niyato, K. Yang, and J. An, "Auction-based time scheduling for backscatter-aided RF-powered cognitive radio networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 3, pp. 1684–1697, 2019.

[20] S. Singh, H. S. Dhillon, and J. G. Andrews, "Offloading in heterogeneous networks: Modeling, analysis, and design insights," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 2484–2497, 2013.

[21] W. Bao and B. Liang, "Stochastic analysis of uplink interference in two-tier femtocell networks: Open versus closed access," *IEEE Trans. Wireless Commun.*, vol. 14, no. 11, pp. 6200–6215, 2015.

[22] Y. Liu, Z. Qin, M. Elkashlan, A. Nallanathan, and J. A. McCann, "Non-orthogonal multiple access in large-scale heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2667–2680, Dec. 2017.

[23] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A tractable approach to coverage and rate in cellular networks," *IEEE Trans. Commun.*, vol. 59, no. 11, pp. 3122–3134, 2011.

[24] H.-S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Heterogeneous cellular networks with flexible cell association: A comprehensive downlink sinr analysis," *IEEE Trans. Wireless Commun.*, vol. 11, no. 10, pp. 3484–3495, 2012.

[25] X. Ge, J. Yang, H. Gharavi, and Y. Sun, "Energy efficiency challenges of 5G small cell networks," *IEEE Commun. Mag.*, vol. 55, no. 5, pp. 184–191, 2017.

[26] H. Zhang, B. Wang, C. Jiang, K. Long, A. Nallanathan, V. C. Leung, and H. V. Poor, "Energy efficient dynamic resource optimization in NOMA system," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 5671–5683, 2018.

[27] Y. Zhang, J. An, K. Yang, X. Gao, and J. Wu, "Energy-efficient user scheduling and power control for multi-cell OFDMA networks based on channel distribution information," *IEEE Trans. Signal Process.*, vol. 66, no. 22, pp. 5848–5861, 2018.

[28] S. N. Chiu, D. Stoyan, W. S. Kendall, and J. Mecke, *Stochastic geometry and its applications*. John Wiley & Sons, 2013.

[29] V.-D. Nguyen, T. Q. Duong, H. D. Tuan, O.-S. Shin, and H. V. Poor, "Spectral and energy efficiencies in full-duplex wireless information and power transfer," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2220–2233, 2017.

[30] V.-D. Nguyen, H. D. Tuan, T. Q. Duong, O.-S. Shin, and H. V. Poor, "Joint fractional time allocation and beamforming for downlink multiuser MISO systems," *IEEE Commun. Lett.*, vol. 21, no. 12, pp. 2650–2653, 2017.

[31] B. R. Marks and G. P. Wright, "A general inner approximation algorithm for nonconvex mathematical programs," *Operations Research*, vol. 26, no. 4, pp. 681–683, 1978.

[32] F. Fang, H. Zhang, J. Cheng, and V. C. Leung, "Energy-efficient resource allocation for downlink non-orthogonal multiple access network," *IEEE Trans. Commun.*, vol. 64, no. 9, pp. 3722–3732, 2016.

[33] M. R. Zamani, M. Eslami, M. Khorramizadeh, and Z. Ding, "Energy-efficient power allocation for noma with imperfect csi," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 1009–1013, 2019.

[34] G. S. Prabhu and P. M. Shankar, "Simulation of flat fading using MATLAB for classroom instruction," *IEEE Trans. Edu.*, vol. 45, no. 1, pp. 19–25, 2002.

[35] H. Tuy, T. Hoang, T. Hoang, V.-n. Mathématicien, T. Hoang, and V. Mathematician, *Convex analysis and global optimization*. Springer, 1998.